School of Electronic Engineering
and Computer Science

Final Report

Programme of study:
BEng COMPUTER SYSTEMS ENGINEERING

Project Title:

EXAMINE THE EFFECTS OF CLIMATE CHANGE IN AN ECOLOGICAL NETWORK USING MACHINE LEARNING

Supervisor:
DR ATHEN MA

Student Name:
SHENIKA REDDY

Final Year
Undergraduate Project 2023/24

Queen Mary
University of London

Date: 29 April 2024

Abstract

Climate change has been a natural phenomenon for millions of years due to factors such as volcanic eruptions, changes in the Earth's orbit and shifts in the Earth's crust. However, since the Industrial Revolution climate change has accelerated greatly, largely due to human activities including the burning of fossil fuels, deforestation, and agriculture. The biggest effect due to climate change is the changing environments for ecosystems all around the world. These shifts can alter the phenology of the species in the ecosystem, as well as changing the flow of energy and interactions between species. Food-webs of an ecological network can be very elaborate, but any small change could result in much larger causal effects on the entire ecosystem. This is the reason why it is important to protect these ecosystems from changes in climate such as global warming, rise in sea levels, increase in carbon dioxide, and changed weather patterns. In order to protect these species from any harmful effects, we first need to understand how the ecosystems are affected.

As ecosystems are made up of a very large number of species all interconnected in different ways, they are complex to analyse. This project aimed to represent this intricate web of interconnected species and their relationships as a graph network, using this to apply techniques used in other network analysis to ecological networks in order to deduce the effects climate change can have on the interactions between species and, therefore, the impacts on the overall stability of the network as a whole.

Contents

Chapter 1: **Introduction**

1.1 Background

It is well known that climate change has a large impact on all different types of ecosystems across the world. It is also known that, in response to climate change, organisms or species undergo various adaptations in order to continue surviving in their new environment. These transformations may include altering their physiology, growth, or reproduction, as well as migrating to habitats with conditions more suited to them. (Peñuelas, J. et al., 2013). This would change the structure of food-webs from these different habitats or ecosystems. One of the greatest effects of climate change is global warming, this worldwide change in temperature has the potential to 'strongly stabilize or destabilize populations and food-webs by changing the interaction strengths between predators and prey' (Rall, B.C. et al., 2010). Due to the fact that the strength of interactions between species in a food-web depend greatly on temperature (Gilbert, B. et al., 2014), any changes in temperature have the potential to substantially shift the strength and stability of the feeding relationships these species have. Small changes in precipitation and nutrient cycling, along with temperature, can also alter the processes of an ecosystem and the structure of the community (Emmerson, M. et al., 2004). This makes it apparent that the relations between species in the food-web of ecological systems can experience changes due to global warming, among other effects of climate change.

Current techniques to explore the effects of climate change on ecological networks include a mathematical approach (as shown in Emmerson, M. et al., 2004, 2005) where the patterns of interactions in the food-web is analysed and formulae are derived in order to model effects. Another approach is empirically investigating the potential impacts of changes in predators and the influence this has on the entire ecosystem (Schmitz, O.J. et al., 2007). While there have been many studies and experiments carried out to identify the effects of climate change, mainly considering global warming, on the stability of species at different levels of the food-web there is a need for further investigation into how these changes will impact interactions between populations as a whole (de Sassi, C. et al., 2012).

1.2 Problem Statement

At the moment, most methods used to investigate the effects of climate change on an ecological network involve taking samples from these ecosystems and performing experiments, then evaluating the results empirically. While this does give a small insight into the effects on particular aspects of the network, it does not necessarily give a picture of the network as a whole as there are many factors involved. Therefore, manually investigating an entire ecological network is not feasible. This creates the need for a more suitable way to represent a complete ecosystem in order to model the effects of potential changes. While other graph modelling and machine learning techniques have been used to analyse an

ecological network, it raises a question of whether network analysis techniques based on that of other graph networks, for example social networks, can be used.

1.3 Aim

The aim of this project is to represent the food-webs of an ecological network as a graph network and to use graph data science based network analysis, along with machine learning algorithms, to analyse the ecosystem and determine the changes that could occur regarding the species themselves, as well as the interactions of species due to climate change. With the results of this analysis, this project seeks to utilize the data to determine the outcome of potential changes and present this so the potential influence on the species in these ecosystems is evident.

1.4 Objectives

To investigate how climate change affects ecosystems.

To investigate current methods for analysing an ecological network.

To identify any drawbacks in these approaches and consider how they may be improved.

To build a graph network model of different ecological networks.

To investigate machine learning Link Prediction algorithms and apply this to ecological networks.

1.5 Research Questions

The results of the network analysis will be used to answer the following questions.

1. What network analysis techniques are most suitable to be applied to an ecological network?

2. How can these techniques be used to investigate effects of climate change?

3. How can the potential impacts of climate change be identified using network analysis techniques?

Chapter 2: **Literature Review**

2.1 Ecosystems and Climate Change

2.1.1 Interactions and Effects of Climate Change

All ecosystems are made up of a network of interactions between species. These interactions are asymmetrical, meaning the interactions are unidirectional, so one species may be dependent on another, but this is not necessarily true the other way around (Griffith, G.P, et al., 2018). Complex ecological networks appear to have a few key species which the network depends on more than the rest, with the other species having more limited interactions. This means that losing a less important species would not have a large impact on the network, but these types of networks are incredibly sensitive to losing one of the key species (Bascompte, J., et al., 2006, Griffith G.P. et al., 2018). The stability of an ecosystem depends on the distribution of trophic links and interaction strength, which itself depends on factors such as predator and prey body sizes, metabolic rate of predators and external conditions (Brierly, A.S. and Kingsford M.J., 2009, Brose, U., et al. 2006, Emmerson, M.C. and Raffaelli, D., 2004, Rall, B.C., et al., 2010, Wootton, J.T. and Emmerson, M.C, 2005). Rather than taking each body mass as an individual measure, the body-mass ratio of predator to prey is used to determine the interaction strength and, through this, the ecosystem stability (Brose, U., et al., 2006, Wootton, J.T. and Emmerson, M.C, 2005). Specifically, as discovered by Wootton, J.T. and Emmerson, M.C (2005) the per-capita interaction strength increased in relation to the ratio of prey to predator weight, albeit with a few inconsistencies. An ecosystems stability can also be considered its resistance or resilience to perturbations. A resistant system would be able to tolerate a large range of conditions, while resilient systems are able to revert back to their original state once the disturbance has passed (Brierley, A.S. and Kingsford, M.J., 2009). The biggest impact of climate change on the conditions of an ecosystem is global warming and there have been countless studies on the effects it has on the interactions and stability of different ecosystems (Brierly, A.S. and Kingsford M.J., 2009, Griffith G.P. et al., 2018, Rall, B.C., et al., 2010). In any ecosystem these changes in temperature can cause changes in growth of the species, leading to changes in the community size structure and their body-mass ratio, as well as affecting muscle development which could cause a change in the movement capabilities of predators and prey. In marine ecosystems, there is a rise in sea level and carbon dioxide (CO_2) content, as well as acidity. The increase in CO_2 causes the pH level of the ocean to decrease, making it more acidic. Warmer waters are less soluble to oxygen (O_2) and are able to accommodate less CO_2 than colder waters, which means they are more prone to acidification, as well as becoming highly saturated more quickly. When the oceans are saturated, they absorb less CO_2, which means more stays in the atmosphere and this causes even more warming, creating a cycle which increases temperature even more rapidly. All of these effects cause a change in phenology and behaviour of species, which impacts the ecosystem that they are a part of. Species that rely on calcium carbonate for their structure or skeletons suffer due to acidity as there is less carbonate available, and predators will be impacted as species which rely

on oxygen would move to respond to varying oxygen zones. Increasing acidity also could interfere with ion exchange, which would decrease metabolism and further decrease the window of thermal tolerance of a species (Brierley, A.S. and Kingsford, M.J., 2009).

Looking at the impacts of climate change, as seen in the experiment conducted by Rall, B.C., et al. (2010) the metabolic rates of predators increased with temperature, and the ratio of ingestion to metabolism decreased due to global warming. This could suggest that there is less available prey, but considering the increase in metabolic rate the predators could also be expending the energy gained from food more quickly. Along with the possibility that there is less prey available, or that the nutritional content of prey may not be as valuable (Wilder, S.M., 2019), this decrease in the ratio would have a larger impact on an ecosystem's stability the more the temperature increases. The overall statistical analysis showed that the rates of metabolism and ingestion of these species depend on its body mass and temperature. The specific effects of climate change on marine ecosystems are altered weather patterns, rising sea levels, oceans being more acidic and the circulation of oceans changing. This causes a change in the range of species, as well as shifts in ecosystem regime. (Brierly, A.S. and Kingsford, M.J., 2009). The experiment by Griffith, G.P., et al. (2018) investigates how any differences in geographical overlap of species could affect an ecosystem. A decrease in geographical overlap suggests that the species are more sensitive to environmental changes, and the reason for this change in overlap is mainly ocean warming. Individual species that have a high geographical overlap with other species in the network can be considered key species, this means the robustness of the network is more reliant on these species. So, if climate change were to alter the overlap of these species with the rest of the ecosystem it may have a detrimental effect. It was found that there was a decrease in interactions of a pelagic predacious species while demersal predators became more structurally important, and the cause of these changes was change in ocean temperature and the resulting adjustment to preferred temperature ranges.

2.1.2 Previous Methods

In most investigations the general method seems to be taking a few sample species from the ecosystem being studied, selecting slightly different sizes or types of species in order to build a rough representation of the real ecosystem in nature. Experiments are then conducted using these species and the results recorded and then analysed using a range of mathematical or modelling approaches.

The experiment by Rall, B.C., et al, (2010) was carried out using different sized terrestrial arthropods, in this case beetles and spiders, and their prey (also of varying sizes). The O_2 consumption of these species was measured at different temperatures, and this was converted into energetic equivalents of metabolism to find how metabolism varied with temperature. Then a subset of these predators was taken, and their ingestion measured at different temperatures with differing prey body masses. The results were used to calculate the ingestion rate, dimensionless energy efficiency, and long-term interaction strength as well as generating regression models.

In the experiment conducted by Griffith, G.P., et al. (2018), the interactions between species in the ecosystem are considered as a network in order to quantify the strength of the network by the proportion of geographical distribution that overlaps for the species under investigation. The proportion of overlaps of an individual species with others in the network is also considered, as this provides an insight into that species' contribution to the robustness of the ecosystem. The frequency distribution of the networks tends to follow an exponential or power law distribution, so in this experiment the species with many connections are considered key species. The results found were used to plot the Pearson's correlation of the species overlap with sea level and surface temperature for both the pelagic and demersal species used in the experiment.

Mukherjee, J., et al., (2015) conducted an experiment to investigate the robustness of an estuarine system using ecological network analysis. A system is considered robust if it is able to maintain certain qualities or return to original functions in response to a certain level of perturbations. The 'Ecopath with Ecosim' software (EwE, https://ecopath.org/) was used to represent networks of carbon exchanges in the Mdloti estuary as a graph. This software allows for mass-balance modelling, time-dynamic simulation, and network analysis. Like all network analysis, the accuracy and completeness of the model depends on the input parameters, however, when making estimations on missing parameters, there are certain assumptions made which may not be accurate for all ecosystems or time periods. The network analysis tool does provide the ability to analyse a variety of indicators of the structure, function, and maturity of the ecosystem but this can be difficult to then interpret, and the results may also be dependent on ecosystem context (Christensen, V. and Walters, C.J., 2004). This experiment by Mukherjee, J., et al., (2015) included changing the biomass of producers and prey, then observing changes in the structure of pathways in the graph.

Most similar to the aims of this project is the study by Araujo, M.B., and Luoto, M., (2007) which investigates the relationship between species and their climate, as well as host plants, using Generalised Additive Modelling (GAM). GAM is a form of statistical machine learning which is used to analyse complex, non-linear relationships such as those between species in an ecological network. The study used multiple variable selection strategies to the GAM in order to test hypotheses and predict changes based on biotic interactions due to climate change.

As summarised by Christin, S., Hervet, É. and Lecomte, N., (2019) deep learning and machine learning have been used in various different ecological studies. These tasks include identifying (Barre, P., et. al. 2017, Knight, E., et al. 2017), and classifying (Li, K., et. al. 2017, Sevilla, A., Bessonne, L., and Glotin, H. 2017) species, studying animal behaviour and interactions (Browning, E. et. al, 2017), modelling ecosystems and monitoring populations (Kroodsma, D. A., et al. 2018). All of which can be used to manage ecosystems in order to assist in conservation as seen in the papers by Di Minin, E., et. al, (2018), and Mohanty, S.P., Hughes, D.P., and Salathé, M., (2016). The study by Drake, J. M., Randin, C., and Guisan, A. (2006) shows how support vector machines can be used to model specific types of ecological networks, while Jeong, K.S., et al (2006) used recurrent artificial neural networks to model the dynamics of phytoplankton. In addition to this, Lek, S., et. al. (1996) show how neural networks are better suited to modelling ecological networks as opposed to using multiple regression. These

studies show how machine learning can be used in a wide range of applications related to investigating ecological networks. Specifically, the wide capabilities of Link Prediction mean that this technique has the potential to be used to uncover connections in an ecological network that may not be immediately obvious, identify anomalies that could suggest changes in the ecosystem or predict any future links between species (Arrar, D., Kamel, N. and Lakhfif, A., 2024).

2.1.3 Dataset

The dataset that has been used is the Gateway Global Database of Traits and Food Web Architecture by Brose, U., (2019). The paper by Brose, U., et al. (2019) titled 'Predator traits determine food-web architecture across ecosystems' outlines an overview of this dataset, which is summarised here.

The dataset is made up of 290 food-webs across the world, with types ranging from freshwater and marine to terrestrial above and below ground. The purpose of the dataset was to develop predator-trait models in order to predict average body-mass ratios, as the stability and functioning of an ecosystem depends on the pattern of body-size architecture in the ecosystem. A food-web with a high predator-prey body-mass ratio has 'weak interactions with slow dynamics that stabilise communities against perturbations', so the aim of this dataset was to find the predator groups with high body-mass ratios in complex food-webs. Examples of possible perturbations include extinction or invasion of species, or other environmental changes due to human effects including global warming. However, different types of ecosystems will have different reactions to changes, so when making any comparisons it is important to compare the same type. As the food-webs are intricate and therefore difficult to describe, the predator traits and interaction strengths are used to find the species that control stability of the community without having to quantify the whole interaction. In order to predict the species with high body-mass ratios with more accuracy, the metabolic and movement types are also included. Analysis of the dataset shows that the body-mass ratio scaling could be dependent on ecosystem type, predator metabolic group, predator movement type or resource supply and feeding behaviour. In accordance with this, aquatic communities seem to have higher average body-mass ratios than terrestrial communities. Overall, body-mass ratio seems to increase with predator and prey mass so larger predators and prey would be characterized by a higher body-mass ratio. In addition, the metabolic and movement types of predators were found to be the most adequate co-variables to consider. A high accuracy was found in predicting the predators with the highest body-mass ratio using the model of only predator traits, with ecosystem type as the independent variable, proving that if a predator is affected by changes in the environment, the whole ecosystem will be affected. This shows that information on food-web structure and prey traits is not necessarily needed, so using the predator-trait model provides a more general solution to assess an entire ecosystem without looking at all or almost all of the specific links and prey traits in the network.

2.2 Graph Networks and Machine Learning

2.2.1 Structure and Applications

The Neo4j Graph Data Science Library has been used to build a graph of the ecological networks; this uses a Property Graph model where the edges of the graph are considered Nodes, and the Vertices are referred to as Relationships. Nodes are assigned labels which allows them to be grouped and Relationships are assigned a type to identify them by. Each of these can also be assigned properties, which are key/value pairs that provide more information about the node or relationship (graphacademy Neo4j Fundamentals, n.d.). The library provides common graph algorithms, as well as machine learning pipelines which are able to use the data stored in the Property Graph to train supervised models and make predictions on the data.

The graph algorithms and machine learning pipelines have been applied to perform ecological network analysis. The article by Fath, B.D., et al. (2007) outlines the steps of constructing ecological networks and provides a framework for how these networks can be used to investigate an ecosystem. Ecological network analysis is described as ‘a systems-oriented methodology to analyze within system interactions used to identify holistic properties that are otherwise not evident from the direct observations’, this means that applying ecological network analysis to examine a network would be able to give an insight into properties of the ecosystem that are embedded more deeply within it, which may not be as practical to identify from observations alone. There is a trade-off in this type of analysis between the amount of data included and the accuracy of the model, as using less species may mean that the larger ecological context is not considered but including too much data would cause the analysis to be very computationally heavy. Therefore, an important aspect of building an ecological network is ensuring that the data needed to quantify all types of species and their interactions should be used.

2.2.2 Algorithms

Centrality, similarity, and link prediction algorithms have been chosen to carry out the analysis of the network. When deciding on the most suitable algorithms to use, the purpose of using it, what results it would yield, and how suitable it would be for the type of graph all had to be considered. Neo4j provides a wide range of centrality algorithms including, but not limited to; Closeness (Fig. 1), Betweenness (Fig. 2), Degree, and Eigenvector Centrality. Each centrality algorithm uses a different measure to determine whether a node is central. Degree centrality and Eigenvector centrality both consider nodes that are highly connected more central, with Degree Centrality only taking into account the number of connections of the nodes while Eigenvector Centrality also considers the weight. In contrast, Closeness and Betweenness Centrality both define the importance in terms of the shortest paths between nodes (Neo4j Graph Data Science Library Manual v2.6, 2023). Degree Centrality works by calculating the sum of relationships, either incoming, outgoing or both, from a node. Eigenvector centrality calculates a weighted sum of the centralities for all nodes connected to the node under investigation. When using closeness centrality, the shortest average path to all other nodes is calculated. Similarly, betweenness centrality

calculates the shortest average path and identifies the nodes that exist on the highest number of shortest paths. (Xambó, 2024)

Closeness Centrality

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j) \right]^{-1}$$

Figure 1. Reciprocal of the sum of the length of shortest paths between the node, i, and all other nodes in the graph.

Betweenness Centrality

$$C_B(i) = \sum_{j \neq i \neq k} g_{jk}(i) / g_{jk}$$

Figure 2. The sum of the shortest paths between two nodes, j and k, that pass through a certain node, i, divided by the total number of shortest paths.

(Xambó, 2024)

In addition, there are two similarity algorithms available, these include Node Similarity and K-Nearest Neighbours. The Node Similarity algorithm can compute either the Jaccard Similarity Score or the Overlap coefficient, these indicate whether two nodes share many of the same neighbours. If the calculated score is high, then the nodes can be considered similar. The K-Nearest Neighbours algorithm compares the properties of each node and computes the distance for all pairs of nodes, then creates relationships based on this between each node and the k neighbours closest to it. It restructures the relationships of the graph so that nodes with the most similar properties are closest to each other. Metrics that can be used to measure similarity for scalar numbers include one divided by one plus the absolute difference of the property values. While it is the Overlap coefficient or the Jaccard similarity for a list of integers, and the Cosine similarity, Pearson correlation score or Euclidean similarity for a list of floating-point numbers. The sampling is done by first picking k random neighbours for each node, in either a Uniform where the neighbours are uniformly chosen at random, or a Random Walk approach where a depth based random walk for k unique nodes are the initial random neighbours and after a defined number of steps, if the number of neighbours has not been visited, a Uniform approach is used to fill the remaining neighbours (Neo4j Graph Data Science Library Manual v2.6, 2023).

Link Prediction algorithms are used to analyse complex networks to find missing links or predict new potential links, the most common use is in recommendation systems, but it can also be used to observe the future changes and interactions in a network (Mutlu, E.C., et al., 2020), which is the purpose it has been used for in this project. The current graph would have to be split into a test set and training set, with the training set to be used when developing the pipeline. The algorithm should not be given the test set until it is complete, as the purpose of the test set is to observe the accuracy of the model. Link Prediction is a powerful tool as it finds the result based solely on the network, without predefined heuristics and

assumptions, which should mean it is more accurate (Mutlu, E.C., et al. 2020). The Neo4j library applies a training model to learn where relationships should exist between pairs of nodes in a graph, and labels nodes as either adjacent or not adjacent through logistic regression. It includes the whole process from feature extraction to the actual link prediction (Neo4j Graph Data Science Library Manual v2.6, 2023).

Chapter 3: **Analysis and Design**

3.1 Selected Algorithms

There are various types of centrality and similarity algorithms available, so the most suitable ones had to be selected.

The Neo4j Degree Centrality, K-Nearest Neighbours and Link Prediction algorithms have been applied as they have been deemed most useful to build an overall picture of the resulting ecosystem due to the effects of climate change. Specific information about the algorithms has come from the Neo4j Graph Data Science Library Manual v2.6, 2023. When taking into account the structure of an ecological network, the species with the most connections generally are considered the most important, or 'key' species (Griffith, G.P., et al., 2018). As there is no edge weight to consider, Degree Centrality is best to be used to find which nodes have the most edges; and this can be used to identify key species in the network. K-Nearest Neighbours is very comprehensive and able to find the similarity between lots of different types of pairs of data. This is beneficial as a larger range of data can be used to find nodes that have the greatest similarity, which could increase accuracy. It also considers both the distance and properties of the graph, meaning that the similarity can be based on specific measures, such as the body mass ratio of consumers and resources. Therefore, it is the more beneficial algorithm to find the strength of interactions so the distribution of trophic links can be discovered. The purpose of implementing Link Prediction is to give an overview of the differences in interactions between species across the network which could brought about by climate change, and this is done by observing the current structure of the ecosystem to determine links which may occur in the future. It can also be used to predict ay potential new links that may develop due to the removal of a species.

Neo4j also provides different modes of execution for each algorithm, including stream, stats, mutate and write mode. Each execution mode provides a different method of running the algorithms as well as returning different results. In order to run an algorithm on the graph, a projection must first be created. Stream mode simply carries out the operation then returns the results of the algorithm as Cypher result rows, such as returning the node ID and value calculated. Rather than returning the actual resulting values, stats mode returns the statistical results for the computation of the algorithm as a single row. In mutate mode, the results of the algorithm are written back to the projected graph and only the statistics from the computation are returned in a similar way to the stats mode. Finally, the write mode writes the algorithm results directly back to the Neo4j database itself rather than just to the projection.

3.1.1 Degree Centrality

This is the simplest among the algorithms used. The number of incoming or outgoing relationships from a node are measured and then these values are returned. By looking at the returned numbers the most popular nodes can be identified as these would have the highest number of incoming and outgoing relationships. It can be used in directed or undirected graphs, as well as with

unweighted or weighted relationships. However, if applied on a heterogeneous graph it will treat it as though it is a homogeneous graph based on the traits of the algorithm. In stream mode, this centrality algorithm returns the node ID and the centrality score.

3.1.2 K-Nearest Neighbours

A distance function selected depending on the properties of the node being considered is used to calculate the distance value for all pairs of nodes and create new relationships between each node and the k neighbours that are closest to it. Initially, a random set of neighbours is selected and verified, then this is refined through any number of iterations up to the maximum specified in the configuration. The neighbours are chosen at random using one of two methods, either uniformly or from a depth biased random walk and in each iteration only a sample of the possible neighbours is compared, with the sample rate also needing to be specified.

The different distance functions that are available depend on if the property is a scalar number, a list of integers or a list of floating-point numbers. The property that has been compared in this project is a list of floating-point numbers so there are three functions that could be used: the Cosine similarity (Fig. 3), the Pearson correlation score (Fig. 4), or the Euclidean similarity (Fig. 5). These metrics take the specified property of each source and target node as inputs, for example their body masses, then compute how similar they are using different techniques.

Cosine similarity

$$cosine(p_s, p_t) = \frac{\sum_i p_s(i) \cdot p_t(i)}{\sqrt{\sum_i p_s(i)^2} \cdot \sqrt{\sum_i p_t(i)^2}}$$

Figure 3. Dot product of the vectors divided by the product of their lengths.

Pearson correlation score

$$pearson(p_s, p_t) = \frac{\sum_i (p_s(i) - \overline{p_s}) \cdot (p_t(i) - \overline{p_t})}{\sqrt{\sum_i (p_s(i) - \overline{p_s})^2} \cdot \sqrt{\sum_i (p_t(i) - \overline{p_t})^2}}$$

Figure 4. Covariance divided by the product of the standard deviations.

Euclidean similarity

$$ED(p_s, p_t) = \sqrt{\sum_i \left(p_s(i) - p_t(i)\right)^2}$$

Figure 5. The root of the sum of the square difference between each pair of elements.

(Neo4j Graph Data Science Library Manual v2.6, 2023)

In stream mode this algorithm returns the nodes it has compared as node 1 and node 2, along with their similarity score.

3.1.3 Node Embedding and Link Prediction

Predicting links in a network require a representation of the graph containing information such as the dimensions and features of the nodes and edges (Grover, A., Leskovec, J., 2016). A node embedding is an algorithm that computes a 'low-dimensional vector representation' (Neo4j Graph Data Science Library Manual v2.6, 2023) of a graph. It generalises and extracts all the important features of the graph while mapping it onto an N-dimensional space (Memgraph, Introduction to Node Embedding, 2021), the results of which can then be used as an input to a Link Prediction algorithm. The node embedding algorithm available through Neo4j which has been used is Node2Vec, this algorithm calculates the representation by sampling the nodes through second order random walks. These random walks traverse randomly chosen edges in the graph. The motivation for using Node2Vec is that, as can be seen in the evaluation by Grover, A. and Leskovec, J., (2016), its performance for Link Prediction surpasses that of other algorithms when compared with heuristic scores and other methods for feature learning. There are four operators which are available for the embedding to be used as a link feature, which is the binary classifier used to predict the probability of a link existing or not. These operators are L2 (Fig. 6), Hadamard (Fig. 7), Cosine (Fig. 3) and Same Category, which is where the category of two nodes are compared and the feature is 1 if they are the same, otherwise it is 0. The Hadamard and Cosine operators were both applied to the embeddings of my dataset in order to combine them into link features which were then used to predict links.

L2

$$f = [(s_1 - t_1)^2, (s_2 - t_2)^2, ..., (s_d - t_d)^2]$$

Figure 6. Set of the sum of squared distances between corresponding elements of two vectors.

Hadamard

$$f = [s_1 * t_1, s_2 * t_2, ..., s_d * t_d]$$

Figure 7. Element-wise multiplication of two vectors.

(Neo4j Graph Data Science Library Manual v2.6, 2023)

There are various steps involved in implementing a Link Prediction pipeline in Neo4j, made up of three phases. These include deriving the feature, training, and test set, observing relationships in the feature set, adding new properties to the graph, and finally, using the train and test sets to train a Link Prediction pipeline. First, the pipeline had to be configured by creating a pipeline, adding node properties and link features, and configuring the split of the relationships into the different sets. In order to split the data, the k-fold cross-validation algorithm is used. This is where the training dataset is randomly split into k equal-sized subsets, with one-fold per iteration being used as the validation set and the rest

being used as effective training sets. Next, the pipeline needs to be trained and evaluated against the validation set. The model with the highest metric is the best option, so this is used to retrain the model on the whole set. Then finally the trained model can be applied for prediction, with the algorithm returning each node and their probability of having a link when it is run in stream mode.

3.2 Requirements

There are a number of key requirements which needed to be met in order for the network analysis to be as accurate as possible. The requirements ensured that the most constructive final results were produced.

Data must be checked for missing or invalid values, and these should be removed.

Network graphs must be structured to connect consumers to resources.

Data required to run each algorithm should be determined based on research.

Nodes in the graph should contain all data required for analysis.

Number of graphs required for thorough analysis should be determined.

3.3 Design

In order to implement the algorithms effectively for different purposes there are certain configuration parameters which can be set to provide the most useful results. To decide what these parameters should be a very small sample of the dataset was taken. The Degree Centrality, K-Nearest Neighbours and Link Prediction algorithms were applied to this subset of the data a few times with different parameters and the results analysed to determine the best parameters to use.

3.3.1 Degree Centrality

The configuration parameters in Degree Centrality are all optional. One of the main purposes is to filter the graph so that the computation only considers certain types of nodes or relationships, this is done by setting the node labels, relationship types or relationship weight property parameters. Other parameters include the concurrency to modify the process of the algorithm concerning the number of threads used to run it, as well as the job ID which can be used to track the progress. In addition to this, there is an orientation parameter to specify the direction of relationships used to compute the degree. As the purpose of calculating the centrality was mainly to find the key species in the network, and due to the fact that there was only one relationship type in this ecological network graph the configuration parameters did not need any changes from the default values.

3.3.2 K-Nearest Neighbours

There is one compulsory configuration parameter for K-Nearest Neighbours, and this is the node properties which will be used in the calculation for the similarity along with the chosen metrics. Other optional configurations include the options included in the Degree Centrality algorithm along with a few others that are specific to the running of this algorithm. The top K specifies the number of neighbours to find, the sample rate, delta threshold and max iterations provide the ability to limit the number of comparisons per node, the value at which to stop running the algorithm and the maximum number of iterations. As well as these, there is the initial sampler parameter which controls the method used to sample the first k neighbours along with a few others with similar types of effects on the function of the algorithm.

The node property which was used is the body mass of consumers and resources.

3.3.3 Node2Vec

All configuration parameters which are a part of the Node2Vec algorithm are optional. These include the same parameters which are available for filtering the graph in the Degree Centrality algorithm and the parameters which allow for monitoring the progress of the algorithm. In addition to these, there are parameters to change the length of a walk and the number of walks per node, the distribution of negative and positive samples, the tendency to stay closer to the start node or return to the last visited node and the option to adjust the method to initialise embeddings as well as the dimension of these embeddings, and a few more. The default value of the embedding dimension is 128, and this was used to calculate the node embeddings for the graphs.

3.3.4 Link Prediction

Implementing the Link Prediction pipeline involves configuring the node properties, link features and the relationship splits. The node properties are the attributes or properties of the graph which can be calculated within the projection in the pipeline, this is where the node embedding was implemented in mutate mode to write the embeddings back to the graph. The link features combine properties of node pairs, representing a relationship between nodes. These link features were then used to train the pipeline. Finally, the relationship splits configure how the graph will be divided into the feature, training, and test set. After these values have been set, the training phase begins and different model candidates are trained using logistic regression, with these automatically being validated using K-Fold Cross Validation to select the best model. Once this has finished running, the score of the best model is returned, and this model can be used to train the whole projection again and return the probability of new possible links.

In the node property step the Node2Vec algorithm was run, and the embeddings added to the graph. These node embeddings were then combined into a link feature using the Hadamard operator and this was used to train the model for prediction.

Chapter 4: **Data Exploration**

4.1 Initial Investigation of Data

Code described in this section is available in Appendix C.

Once the initial background research was complete, a detailed look at the data was needed. To do this the Python Pandas and NumPy libraries were used. After importing the data, any object datatypes were changed to categories to make it easier to explore the data. Then a sample of 10 values at a time was investigated, having run multiple times in order to see and understand the structure of the data without looking through all values, as this dataset is very large.

There appeared to be an abundance of missing data, which is as expected for a dataset of this type. These values were either filled with NaN for object/categorical features and with -999 for numerical features. For the purpose of understanding exactly how many of these values are present, the -999 values were replaced with NaN and a loop was implemented to count how many times this null value was present in each column. The result of this is shown in Figure 8.

```
autoID: 0
link.citation: 0
link.methodology: 156171
interaction.type: 1974
interaction.dimensionality: 18374
interaction.classification: 902
con.taxonomy: 0
con.taxonomy.level: 8025
con.common: 164992
con.lifestage: 83616
con.metabolic.type: 0
con.movement.type: 31
con.size.citation: 174764
con.size.method: 183148
con.length.min.cm.: 222151
con.length.mean.cm.: 205851
con.length.max.cm.: 222151
con.mass.min.g.: 222151
con.mass.mean.g.: 861
con.mass.max.g.: 222151
res.taxonomy: 0
res.taxonomy.level: 29758
res.common: 168573
res.lifestage: 113972
res.metabolic.type: 164
res.movement.type: 1503
res.size.citation: 177507
res.size.method: 183148
res.length.min.cm.: 222151
res.length.mean.cm.: 206029
res.length.max.cm.: 222151
res.mass.min.g.: 222151
res.mass.mean.g.: 19485
res.mass.max.g.: 222151
geographic.location: 0
longitude: 0
latitude: 0
ecosystem.type: 0
study.site: 0
altitude: 193803
depth: 182458
sampling.time: 102562
sampling.start.year: 64245
sampling.end.year: 64245
notes: 191839
foodweb.name: 0
```

Figure 8. List of the number of missing items per column in the dataset

This dataset was used by Brose, U., et al. (2019) in an investigation into using predator, or consumer, traits to make predictions about the ecosystem, and the objective of this project is to explore the distribution of links and interaction strengths of these ecological networks, particularly in terms of body-mass ratio. It was determined from the observations that body mass was most suitable to include in the network analysis. Although there are quite a large number of missing values, each dataset was checked for missing values so that the portion of data being used did not contain missing values. While the size of the consumer could also give an insight into the interactions between species, there is a much larger quantity of the values that are missing so it could not be used. In addition to this, it would also be possible to use consumer metabolic and ingestion rates to investigate the distribution of trophic links, but they have not been used as only the types are included in the dataset, not the values.

As part of the initial investigation and preparation of the data it had to be filtered and split based on the location.

The dataset was filtered by location and food-web using the longitude, so each CSV file contained the ID, consumer and resource taxonomy, consumer and resource mass, interaction type and food-web name. As mentioned previously, the body mass columns could have contained missing values, so a check was put in place to ensure each filtered dataset does not have any missing values. After filtering, another check ensured there is some overlap between species in different food-webs from the same location so these could be aggregated for further analysis. A random sample of data has been taken from the filtered dataset of the food-web L1P1 located in Portugal (Table 1).

autoID	interactionType	conTaxonomy	conMassMean	resTaxonomy	resMassMean	foodwebName
374	herbivorous	Pantopoda	0.0015	Ulva lactuca	0.000228	L1P1
233	predacious	Lipophrys trigloides	0.7	Balanus sp.	0.01	L1P1
94	herbivorous	Ectoprocta	0.01	PhytoP	0.0001	L1P1
388	predacious	Polychaeta	0.013	Balanus sp.	0.01	L1P1
282	predacious	Nassariidae	0.06	Insecta	0.01	L1P1

Table 1. Sample of L1P1 dataset.

It was decided that the focus of this project would be on marine ecosystems as this data was most abundant, with a large number of locations to be able to determine accurate results. The food-webs chosen were L1P1, L1P2, L2P1, CR1P3 and CR2P4 located in Portugal.

Chapter 5: **Implementation**

5.1 Graphs

The graphs that have been constructed for each of the food-webs located in Portugal contain between 61 and 93 nodes with 151 to 437 relationships individually. The L1P1 graph contains the most nodes and relationships while the CR2P4 graph contains the least.

The graphs are structured with the consumers and resources as nodes, and their interactions as relationships. Using the filtered data, each node has the taxonomy of the species, the food-web it belongs to and the body mass as properties, while the links have the interaction ID and type of interaction as properties. As mentioned, these ecological networks all contain overlapping species but are from the same location and therefore represent similar climates.

Consumers are represented by blue nodes and resources by green.

5.1.1 L1P1

This graph contains 93 nodes and 437 relationships.

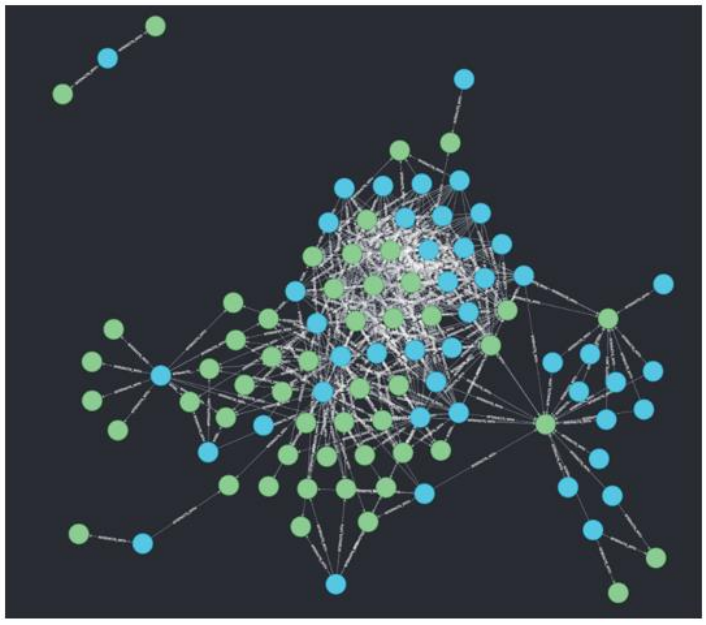


Figure 9. Graph of relationship between consumers and resources in L1P1 food-web, Portugal.

5.1.2 L1P2

This graph contains 71 nodes and 294 relationships.

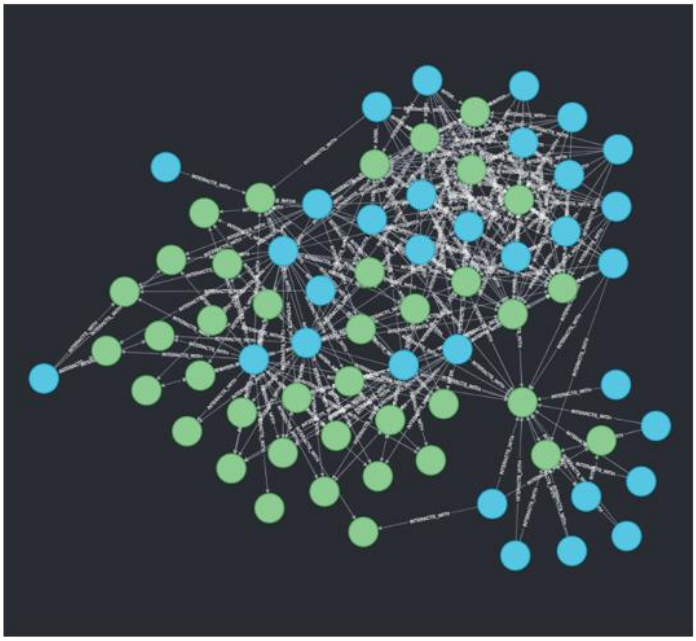


Figure 10. Graph of relationship between consumers and resources in L1P2 food-web, Portugal.

5.1.3 L2P1

This graph contains 77 nodes and 320 relationships.

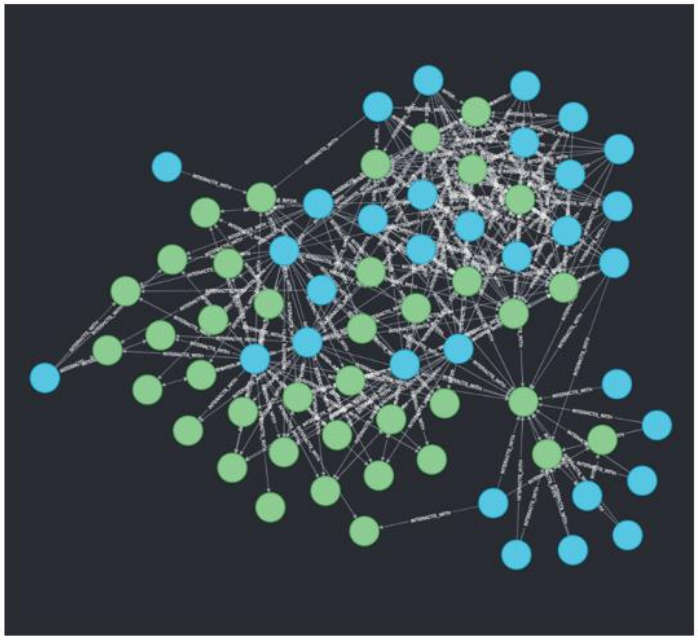


Figure 11. Graph of relationship between consumers and resources in L2P1 food-web, Portugal.

5.1.4 CR1P3

This graph contains 68 nodes and 284 relationships.

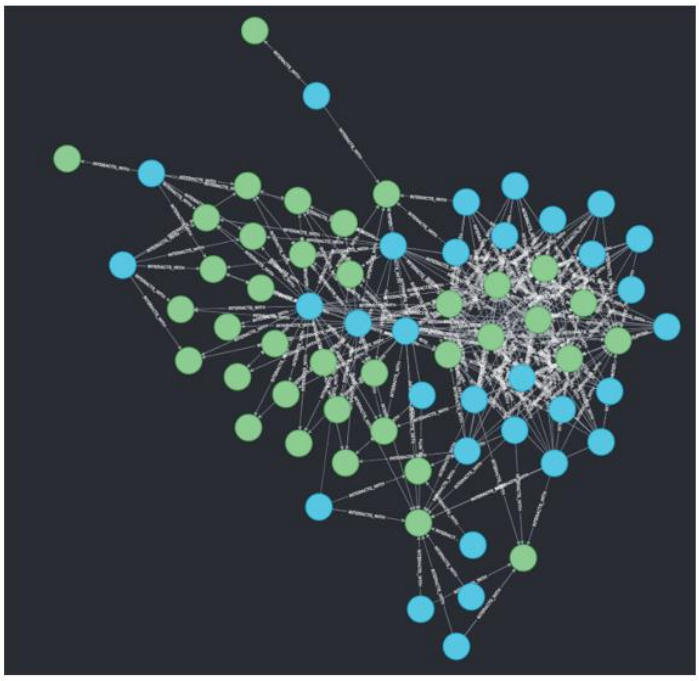


Figure 12. Graph of relationship between consumers and resources in CR1P3 food-web, Portugal.

5.1.5 CR2P4

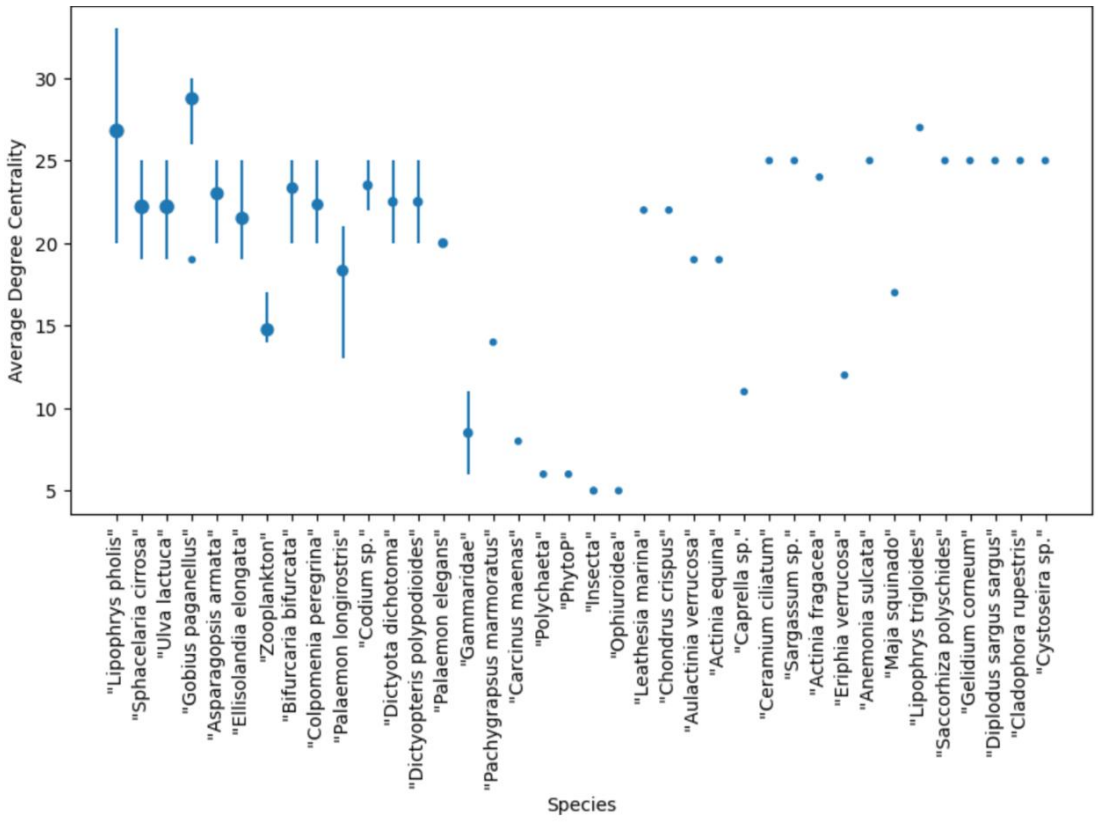
This graph contains 61 and 151 relationships.

Figure 13. Graph of relationship between consumers and resources in CR2P4 food-web, Portugal.

5.2 Centrality

In the Degree Centrality analysis, the 15 nodes that have the highest centrality were considered, as well as recording whether these are consumers or resources. The centrality was calculated by creating an undirected projection of the graph and running the Neo4j Degree Centrality algorithm on this in stream mode, then recording the results. This projection is undirected because it is inconsequential whether a relationship is incoming or outgoing, a node simply needs to be highly connected to other nodes in the graph to be considered important in the structure of the ecological network. The graph plotted shows the average Degree Centrality results of each species, along with error bars to show the range of centrality scores recorded. The node size indicates the number of food-webs in which the species had a high centrality score. For example, it can be seen that Zooplankton was in almost all of the food-webs as the size of the node is large, but it had a lower overall centrality score with a smaller range. In contrast, PhytoP only appears once and has a very low centrality score.

Figure 14. Graph of degree centrality results, including top 15 species across all Portugal food webs.



5.3 Similarity

The K-Nearest Neighbours similarity score was calculated for consumer-resource pairs in each food-web using the body mass of the species. The same projection as used for the centrality was also used to apply the K-Nearest Neighbours algorithm. It was discovered through research that consumers and resources with a greater difference in body mass could be considered to have a weaker interaction strength which makes them more stable against disturbances. So, the

inverse of the similarity was calculated, and this is displayed in the heatmaps (Fig. 15 to 19).

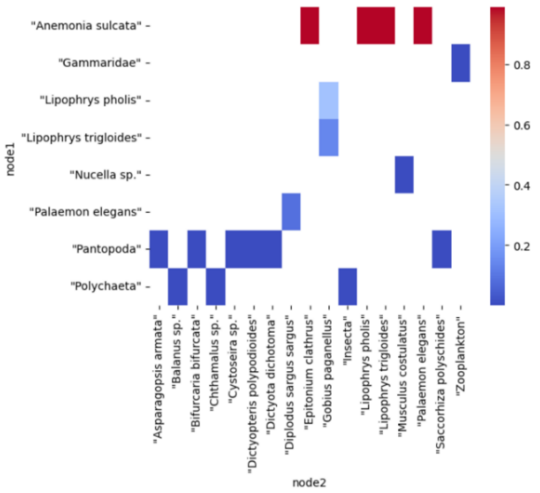


Figure 15. Heatmap of similarity scores for L1P1 food-web.

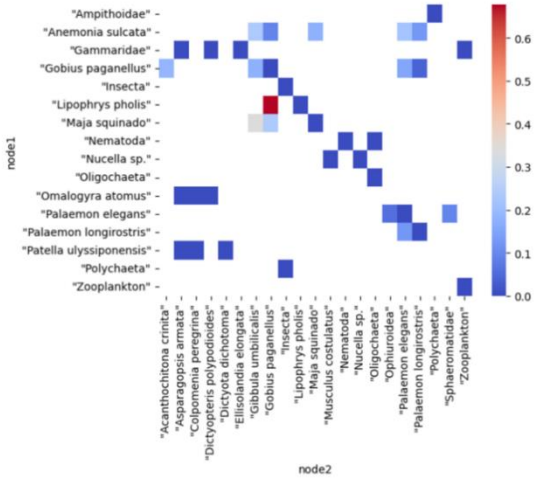


Figure 16. Heatmap of similarity scores for L1P2 food-web.

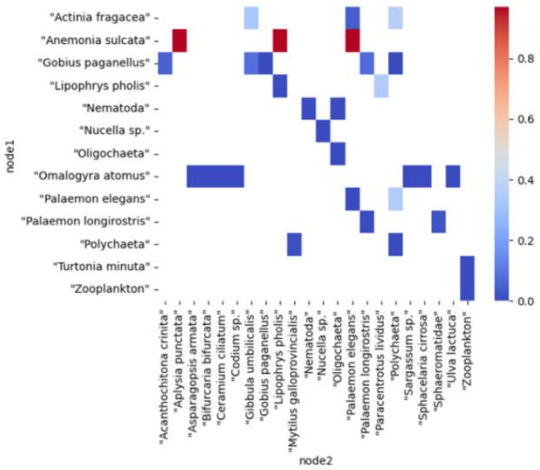


Figure 17. Heatmap of similarity scores for L2P1 food-web.

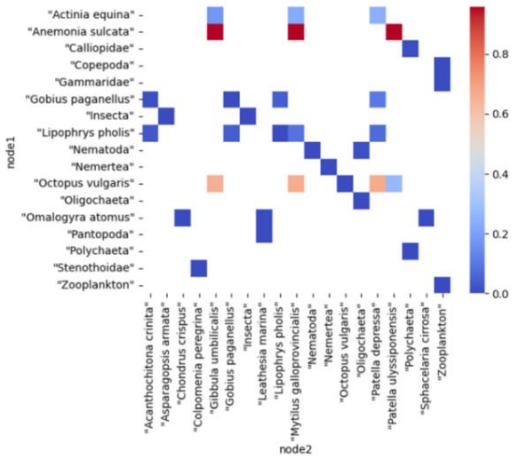


Figure 18. Heatmap of similarity scores for CR1P3 food-web.

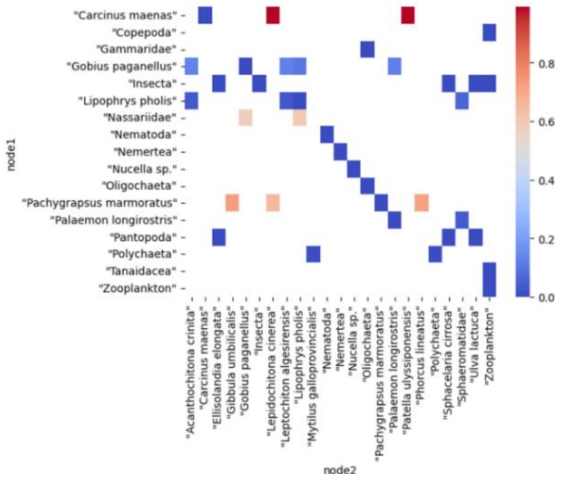


Figure 19. Heatmap of similarity scores for CR2P4 food-web.

5.4 Predictions

The Link Prediction pipeline was implemented on a variety of different graphs using multiple different parameters in order to determine the best model for use in predictions, however in each of these utilisations although the average train score ranged from 70% to 79%, the results did not appear to be valid.

5.4.1 Small Test Graph

First, to observe the operation of the algorithm it was applied to a very small graph containing only 22 nodes and 20 relationships (Fig. 20).

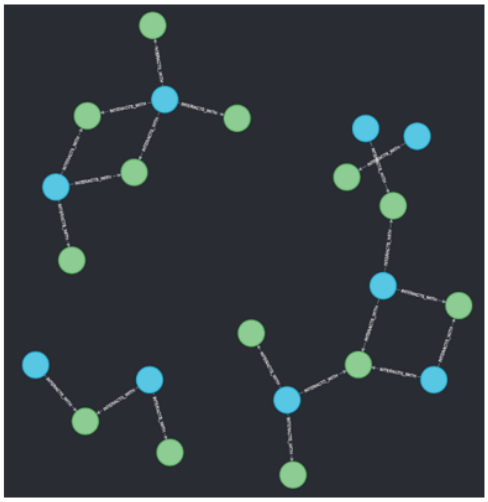


Figure 20. Graph of relationship between consumers and resources in Ythan Estuary, Scotland

The embedding using Node2Vec was calculated, with the embedding dimension set to 42 as the graph was small, so a lower dimension was more suitable to achieve a higher train score. 42 was chosen because it provided a high train score at 93.0% while also giving a larger range of predicted probabilities for links ranging from 0.56 to 0.59. When tested with 56 as the dimension, the average train score reduced to 89.8% and all values for the probability were 0.501. Although the train score did increase to 96.8% with a dimension of 32, all probabilities were the value 0.5. Therefore, although all iterations did not provide definitive results considering the highest probability was below 0.6, an embedding dimension of 42 still appeared to achieve the most fitting results. The node embeddings were then used to calculate the link features using the Hadamard operator, with only the embeddings being considered. While these results may have appeared to be incorrect this could have been due to the fact that the graph was so small. So, despite these results it was determined that the algorithm could perform adequately to be applied to larger graphs to investigate the effects of climate change.

5.4.2 L1P1 Food-Web

The next graph which was analysed was the L1P1 food-web (Fig. 9), which contained 93 nodes and 437 relationships. As this network is larger than the previous, the embedding dimension used with Node2Vec which provided the highest train score was the value 128. At first the Cosine Similarity operator was used to predict links based on the embeddings and the mass of the species under consideration. The average train score was 76.9%, but as before all probabilities returned had a value of 0.5. The same investigation was attempted once again with the Hadamard operator instead of Cosine. While this did increase the train score to 78.7% the probabilities were still 0.5. As the prediction using mass did

not seem to be returning credible results, another pipeline was run using only the embeddings along with the Hadamard operator. This caused the train score to decrease to 71.5% and as previously the predicted probabilities returned were all 0.5. As the prediction still did not appear to be working, one more analysis was attempted.

5.4.3 Portugal Food-Webs

In the case that the L1P1 graph that had been used was still too small to achieve meaningful results, the final analysis was carried out on an aggregate of all Portugal food-webs in consideration. This included L1P1, L1P2, L2P1, CR1P3 and CR2P4. There were 148 nodes along with 821 relationships overall (Fig. 21). Due to species overlap the number of nodes did not increase as much as relationships, in comparison with the individual graphs.

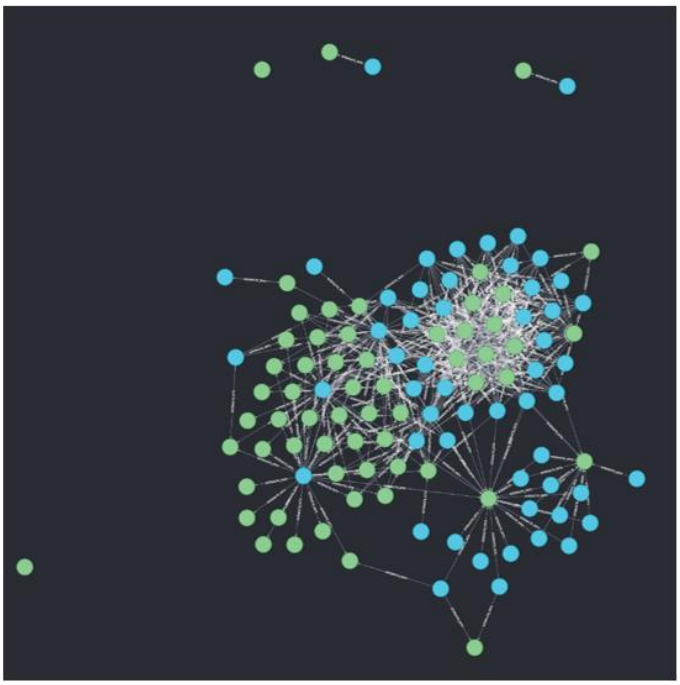


Figure 21. Aggregate graph of relationships between consumers and resources in L1P1, L1P2, L2P1, CR1P3, CR2P4 food-webs from Portugal.

The embedding dimension of 128 which was used in the analysis of L1P1 was also found to be most suitable for this network. The same parameters used in the analysis conducted on L1P1 were applied to this network as well. When using the embedding and mass as link features and the Cosine operator to relate them, the average train score was 61.6% and all probabilities were returned as 0.49. Using the same link features with the Hadamard operator increased the train score to 74.4% but all probabilities were 0.5. Finally, the pipeline was done only including the embedding. This caused the train score to decrease to 65.9% while the probabilities remained at 0.5.

5.4.4 Social Network

As the Link Prediction algorithm did not seem to be working, it was then tested on a social network to find out if there was a problem with the networks or with the application of the algorithm itself. The dataset was retrieved from The Network Data Repository with Interactive Graph Analytics and Visualization (2015) by Ryan A. Rossi and Nesreen K. Ahmed. This graph contained 620 nodes and 2100 edges (Fig. 22), where the nodes represent Facebook pages relating to food, and the edges represent mutual likes between pages.

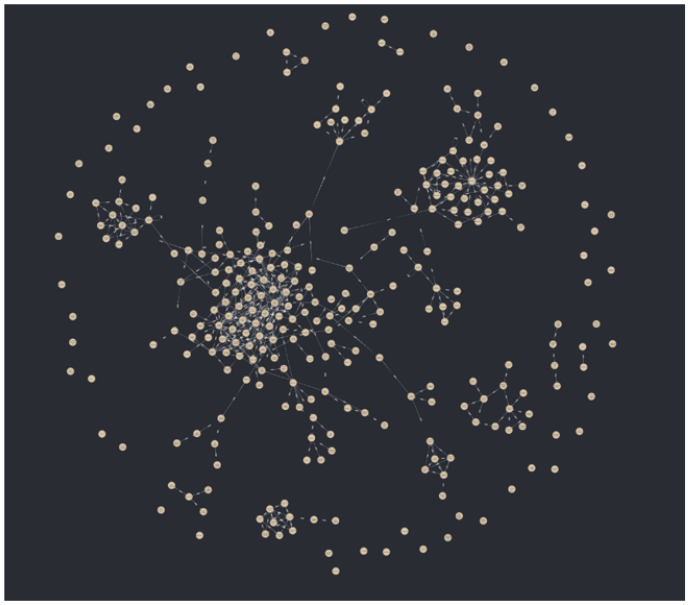


Figure 22. Facebook graph of mutually liked pages

Due to the fact that this graph was larger than the ecological networks, an embedding dimension of 256 returned the best average train score at 87%. This returned results that appeared much more reliable with probabilities for the potential links ranging from 0.79 to 0.84 (Table 2).

Table 2. Sample of link prediction results for Facebook network.

Page1	*Page2*	*probability*
"River Crab Blue Water Inn"	"Grotto"	0.844379991
"Cadillac Bar Houston"	"Gandy Dancer"	0.795113527
"Lillie's Asian Cuisine"	"Ravintola Salve"	0.799802142
"River Crab Blue Water Inn"	"Bubba Gump Shrimp Co."	0.805392837
"Cadillac Bar Kemah"	"RED Sushi and Hibachi Grill"	0.816751098
"Charley's Crab"	"Grotto"	0.806395712
"Chart House"	"Charley's Crab"	0.839017587
"Taco Bell Kuwait"	"Taco Bell Rep Dom"	0.793058197

In order to determine if the node embedding was the cause of these repeated invalid results the accuracy of the node embeddings was calculated had to be examined. This was done by creating a very small graph (Fig. 23) and observing the embeddings that were determined by running the algorithm in only two dimensions (Table 3).

Figure 23. Small symmetrical graph containing two consumers and five resources.

Table 3. Species and the corresponding Node2Vec embedding for small symmetrical graph.

Species	Embedding
"Lutra lutra"	[0.025189561769366264, -1.232606291770935]
"Ardea cinerea"	[-0.0069526489824056625, -1.3101673126220703]
"Anguilla anguilla"	[0.02798759564757347, -1.2642314434051514]
"Pholis gunnellus"	[0.0023159475531429052, -1.192455530166626]
"Zoarces viviparus"	[0.049173761159181595, -1.2176783084869385]
"Pollachuis virens"	[0.025586741045117378, -1.3066909313201904]
"Carcinus maenas"	[0.03607982397079468, -1.4866533279418945]

The similarity score for these embeddings was calculated to observe how closely the nodes are placed in the embedding space. It is expected that the two consumers, ‘Lutra lutra’ and ‘Ardea cinerea’ have the closest embedding. In addition, the resource in the centre, ‘Anguilla anguilla’ may have a slightly different embedding to other nodes, but the remaining four should also have similar embeddings to each other.

Overall, the similarities of all embeddings were around 0.99, which is as expected from observing the results (Table 3). Another consumer was added to the graph to observe any changes in the embeddings, but this still returned embeddings with a similarity of 0.99. When observing the embedding result for the L1P1 network the embedding results also had a similarity of around 0.99. Similarity scores this high indicate that the embeddings are nearly identical, so the structure and node characteristics are not accurately represented meaning it is very difficult to predict new links based on this information.

By looking at the similarities between the embeddings for the Facebook graph, the average similarity was around 0.58. This score shows that the embeddings have a good balance between representing the commonalities between node characteristics while also preserving the overall structure of the graph making these embeddings much more useful to accurately predict links.

The differences in the embeddings are made clear by observing the differences in the heatmaps of similarity scores for the embedding of L1P1 (Fig. 9) and the Facebook graph (Fig. 22). The heatmap of L1P1 (Fig. 24) is mostly red, which shows that most scores are above 0.95 making it abundant how close the embeddings are mapped. It can also be seen that even the lowest scores represented by blue have a similarity of at least 0.8. Whereas the heatmap of the Facebook graph (Fig. 25) is white to a very faint blue, indicating the scores are

roughly between 0.59 and 0.6. Showing how these embeddings are much more balanced and therefore can be used effectively to predict links.

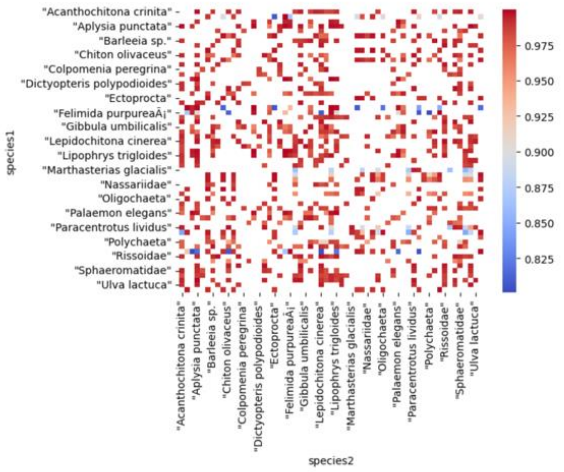


Figure 24. Heatmap of similarity scores for L1P1 node embeddings.

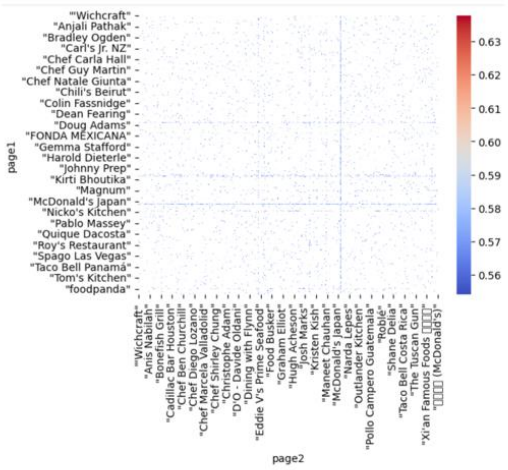


Figure 25. Heatmap of similarity scores for Facebook network node embeddings.

Chapter 6: **Evaluation**

6.1 Identifying Key Species

The Degree Centrality was used to find the most dominant key species in each food-web. The results show that there are certain species which have a high centrality score regardless of which food-web it is a part of, which indicates that these are very important species in this ecosystem. Figure 14 shows the Degree Centrality results. This code used to plot this graph can be found in Appendix C.

It can be seen that the consumer ‘Lipophrys pholis’ has almost the highest average centrality score across all food-webs, as well as the largest range. Consumers having a high centrality means that these species interact with a variety of different types of prey, suggesting they are more resilient to changes in the ecosystem. This is because any removal of resources that a species with a large number of connections relies upon would not have as large an effect as it would on a consumer having the only resource it relies upon removed. This suggests that ‘Lipophrys pholis’ is a particularly resilient consumer as it has a large but high range, while also being present in all food-webs. In relation to this, resources with a high centrality would be important in the ecosystem as many consumers rely on them, so if these resources were to be removed it would have a causal effect on the consumers connected to them. Species such as ‘Ulva lactuca’ and ‘Sphacelaria cirrosa’ which are present in all food-webs and have a fairly high average centrality score are examples of these types of key resources. From the results it is also apparent that there is a balance between key consumers and resources, but resources have a slightly higher average centrality at 20.4 as opposed to the average of 17.5 for consumers. Due to the fact that the difference in averages is not too large, and some species are both consumer and resource, it shows that the stability of these ecosystems depends on the resilience of the consumers and resources when faced with changes in their environment. The key species which are consumers most likely improve the stability of the ecosystem slightly as they are less likely to be affected by the loss of resources. If any resource is removed from the ecosystem due to changes in the climate, this would have a ripple effect on any consumers that rely on it. Consumers that rely on multiple resources would not face very harmful results, while consumers that rely solely on this resource, they would have no source of nutrition and may not survive, or alternatively would have to find another source. This would cause changes in the ecosystem regime, and only the more resilient consumers would survive. Meanwhile, if a key resource is removed the same potential effects would occur, but the effect could be much more significant as there are more consumers that would be affected.

Overall, this information about the key species in the network gives an indication of what the strength of the ecosystem depends on. It can be concluded that the strength of the ecosystem does depend on the number of key consumers and the resilience of resources when faced with change.

6.2 Determining Strength of Interactions

The inverse of the K-Nearest Neighbour similarity results has been displayed in heatmaps where blue represents a low score and red represents a high score (Fig. 15 to 19). The code used to create the heatmaps can be found in Appendix C.

The vertical axis shows node 1, which is the consumer, and the horizontal axis shows node 2, which is the resource. Any relationship that is blue has a small difference in mass and therefore a strong interaction strength, while the relationships that are red have a larger difference in mass meaning these interactions are weaker. The colour that is most abundant across food-webs is blue. So, these results show that in each food-web the majority of the consumers and resources that interact have a more similar mass, indicating a lower body-mass ratio. This can be taken to suggest that if there were to be any differences in these species' environments, it may be more difficult for them to adapt. So, if the similarity in mass can be considered a reliable measure for interaction strength, it can be established that these ecological networks are highly susceptible to changes in climate. However, in reality there are various factors that may affect the strength of interactions, so these results may not be entirely accurate.

6.3 Predicting Relationships

A probability of 0.5 means there is around a 50% chance of the predicted link from occurring. This shows the prediction model is uncertain of whether or not the link should exist, and the prediction can bear comparison to simply making a random guess. The reason for this could be that the model struggles to distinguish between the presence of absence of a link or is unable to use information it is given to confidently predict and provide informed results. It could suggest that the model either requires more refinement or may just be ineffective.

Having tried various different parameters to predict links in the food-webs without many changes in the results, it could be assumed that this method was not effective when used for ecological networks. This was reinforced by the fact that more reliable results were achieved by implementing the same algorithm on a social network. There are a few potential reasons for this, but because the analysis did work on the social network, it could be due to the differences in structure. These differences include the number of nodes, as well as the fact that the social network has very clear established communities with many links between them, while the ecological network is structured without any clear communities and has a much lower number of nodes. The social network is also homogenous while the ecological network graph is heterogeneous. Link Prediction works by inferring the existence of relationships due to the closeness of the node embeddings. This works well with social networks as it can be assumed that users would like content similar to their previous likes. However, in an ecosystem it is not necessarily the case that two consumers that have similar diets would consume all of the same resources, or each other. Therefore, applying this model based solely on nodes having similar neighbours is not the most effective in an ecological network. When calculated, the embedding for the

ecological network was too close together. This could be due to the overlapping diets of species but meant the Link Prediction was not accurate. It may be possible to improve the model by providing more information to the algorithm and making the representation of consumer-resource interactions more similar to real life, but this will not necessarily change the results as this method of Link Prediction is based on the structure of social networks rather than ecological.

Chapter 7: **Conclusion**

7.1 Summary

The aim of this project was to evaluate the response of ecological networks to climate change using graph network analysis techniques including Degree Centrality, K-Nearest Neighbours and Link Prediction. This is of great importance due to the ever-increasing impact changes in climate are having on species in ecosystems that are being forced to adapt to alterations in their environment. These differences pose threats to the structure and stability of the ecosystems and can be dangerous for the species that are a part of these ecosystems.

With the application of Degree Centrality and K-Nearest Neighbours, along with background research, it was found that these algorithms can be applied successfully to an ecological network to identify key species and similarities in relationships between species in order to determine the strength of the ecosystem when faced with changes. The information about key species found that the stability of the ecosystems relied on resources interacting with multiple consumers, as well as key consumers having strong resilience in response to changes. Along with this, when using the ratio of body masses as a measure of interaction strength, it was found that the majority of interactions in the ecosystems studied were strong and therefore less able to stabilise when faced with disturbances. If these measures can be considered reliable the results found suggest changes in climate should be brought under control in order to protect these ecosystems.

Applying the Link Prediction algorithm in the same way as applied to social networks was also discovered to not be reliable when predicting links in ecological networks as the methods used by these algorithms are based on the similarities between nodes, but similarities between species in an ecosystem do not necessarily mean a link exists.

It can be concluded that climate change does have negative impacts on the species within ecological networks, and therefore the structure of these networks as well, because of the difficulty for species to adapt to differences in their environment.

7.2 Further Work

Further work would most certainly include further adjusting the parameters in the Link Prediction algorithm used as well as investigating other network analysis techniques which can be used to predict or observe changes in links between nodes in a network. The accuracy with which these algorithms can model an ecological network should also be taken into consideration. Once these have been found the next step would be attempting to apply these methods to ecological networks in the hope that they return more accurate results.

References

Araújo, M.B. and Luoto, M., (2007). *The importance of biotic interactions for modelling species distributions under climate change.* Global Ecology and Biogeography, 16(6), pp.743-753

Arrar, D., Kamel, N. and Lakhfif, A., (2024). *A comprehensive survey of link prediction methods*. The Journal of Supercomputing, 80(3), pp.3902-3942.

Barré, P., Stöver, B.C., Müller, K.F. and Steinhage, V., (2017). *LeafNet: A computer vision system for automatic plant species identification.* Ecological Informatics, 40, pp.50-56.

Bascompte, J., Jordano, P. and Olesen, J.M., (2006). *Asymmetric coevolutionary networks facilitate biodiversity maintenance*. Science, 312(5772), pp.431-433.

BBC (2013), A brief history of climate change [online]. Available at: https://www.bbc.co.uk/news/science-environment-15874560 [accessed 25/11/2023]

Brierley, A.S. and Kingsford, M.J., (2009). *Impacts of climate change on marine organisms and ecosystems*. Current biology, 19(14), pp.R602-R614.

Brose, U., Archambault, P., Barnes, A.D., Bersier, L.F., Boy, T., Canning-Clode, J., Conti, E., Dias, M., Digel, C., Dissanayake, A. and Flores, A.A., (2019). *Predator traits determine food-web architecture across ecosystems*. Nature ecology & evolution, *3*(6), pp.919-927.

Brose, U., Jonsson, T., Berlow, E.L., Warren, P., Banasek-Richter, C., Bersier, L.F., Blanchard, J.L., Brey, T., Carpenter, S.R., Blandenier, M.F.C. and Cushing, L., (2006). *Consumer–resource body-size relationships in natural food webs*. *Ecology*, *87*(10), pp.2411-2417.

Brose, U. (2019) *GlobAL daTabasE of traits and food Web Architecture* (GATEWAy) v.1.0.(iDiv Data Repository) Available at: https://idata.idiv.de/ddm/Data/ShowData/283?version=3 [accessed 12/10/2023]

Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T. and Freeman, R., (2018). *Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds*. Methods in Ecology and Evolution, 9(3), pp.681-692.

Christensen, V. and Walters, C.J., (2004). *Ecopath with Ecosim: methods, capabilities and limitation*s. Ecological modelling, 172(2-4), pp.109-139.

Christin, S., Hervet, É. and Lecomte, N., (2019). *Applications for deep learning in ecology.* Methods in Ecology and Evolution, 10(10), pp.1632-1644.

Di Minin, E., Fink, C., Tenkanen, H. and Hiippala, T., (2018). *Machine learning for tracking illegal wildlife trade on social media.* Nature ecology & evolution, 2(3), pp.406-407.

Drake, J.M., Randin, C. and Guisan, A., (2006). *Modelling ecological niches with support vector machines.* Journal of applied ecology, 43(3), pp.424-432.

Emmerson, M.C., Bezemer, T.M., Hunter, M., Jones, T.H., Masters, G. and Van Dam, N.M. (2004). *How does global change affect the strength of trophic interactions?*. Basic and Applied Ecology, 5(6), pp.505-514.

Emmerson, M.C., Bezemer, M., Hunter, M.D. and Jones, T.H. (2005). *Global change alters the stability of food webs.* Global Change Biology, 11(3), pp.490-501.

Emmerson, M.C. and Raffaelli, D., (2004). *Predator-prey body size, interaction strength and the stability of a real food web*. Journal of Animal Ecology, pp.399-409.

Fath, B.D., Scharler, U.M., Ulanowicz, R.E. and Hannon, B., (2007). *Ecological network analysis: network construction*. Ecological modelling, 208(1), pp.49-55.

Gilbert, B., Tunney, T.D., McCann, K.S., DeLong, J.P., Vasseur, D.A., Savage, V., Shurin, J.B., Dell, A.I., Barton, B.T., Harley, C.D. and Kharouba, H.M. (2014). *A bioenergetic framework for the temperature dependence of trophic interactions*. Ecology letters, 17(8), pp.902-914.

Griffith, G.P., Strutton, P.G. and Semmens, J.M., (2018). *Climate change alters stability and species potential interactions in a large marine ecosystem*. Global Change Biology, *24*(1), pp.e90-e100.

Grover, A. and Leskovec, J., 2016, *node2vec: Scalable feature learning for networks*. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 855-864).

History.com (2023), *Climate Change History* [online]. Available at: https://www.history.com/topics/natural-disasters-and-environment/history-of-climate-change [accessed 25/11/2023]

Jeong, K.S., Joo, G.J., Kim, H.W., Ha, K. and Recknagel, F., (2001). *Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network*. Ecological Modelling, 146(1-3), pp.115-129.

Knight, E.C., Hannah, K.C., Foley, G.J., Scott, C.D., Brigham, R.M. and Bayne, E., (2017). *Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs.* ACE, 12(2), p.14.

Kroodsma, D.A., Mayorga, J., Hochberg, T., Miller, N.A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T.D., Block, B.A. and Woods, P., (2018). *Tracking the global footprint of fisheries*. Science, 359(6378), pp.904-908.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. and Aulagnier, S., (1996). *Application of neural networks to modelling nonlinear relationships in ecology.* Ecological modelling, 90(1), pp.39-52.

Li, K., Javer, A., Keaveny, E.E. and Brown, A.E., (2017). *Recurrent neural networks with interpretable cells predict and classify worm behaviour*. bioRxiv, p.222208.

Memgraph (2021), *Introduction to Node Embedding* [online]. Available at: Introduction to Node Embedding (memgraph.com) [accessed 21/02/2024]

Met Office (2023), *Causes of climate change* [online]. Available at: https://www.metoffice.gov.uk/weather/climate-change/causes-of-climate-change [accessed 25/11/2023]

Mohanty, S.P., Hughes, D.P. and Salathé, M., (2016). *Using deep learning for image-based plant disease detection*. Frontiers in plant science, 7, p.215232.

Mukherjee, J., Scharler, U.M., Fath, B.D. and Ray, S., (2015). Measuring sensitivity of robustness and network indices for an estuarine food web model under perturbations. Ecological Modelling, 306, pp.160-173.

Mutlu, E.C., Oghaz, T., Rajabi, A. and Garibay, I., (2020). *Review on learning and extracting graph features for link prediction*. Machine Learning and Knowledge Extraction, 2(4), pp.672-704

Neo4j (2023), *Graph Data Science Library Manual v2.6* [online]. Available at: https://neo4j.com/docs/graph-data-science/2.6-preview/ [accessed 25/11/2023]

Peñuelas, J., Sardans, J., Estiarte, M., Ogaya, R., Carnicer, J., Coll, M., Barbeta, A., Rivas-Ubach, A., Llusià, J., Garbulsky, M. and Filella, I. (2013). *Evidence of current impact of climate change on life: a walk from genes to the biosphere.* Global change biology, 19(8), pp.2303-2338.

Rall, B.C., Vucic-Pestic, O.L.I.V.E.R.A., Ehnes, R.B., Emmerson, M. and Brose, U. (2010). *Temperature, predator–prey interaction strength and population stability*. Global Change Biology, 16(8), pp.2145-2157.

Ryan, A.R. and Nesreen, K.A., (2015). *The Network Data Repository with Interactive Graph Analytics and Visualization.* In Proceedings of the AAAI conference on artificial intelligence (Vol. 29, No. 1). Available at: https://networkrepository.com [accessed 08/04/2024]

de Sassi, C., Lewis, O.T. and Tylianakis, J.M. (2012). Plant-mediated and nonadditive effects of two global change drivers on an insect herbivore community. Ecology, 93(8), pp.1892-1901

Schmitz, O.J. (2007). *Predator diversity and trophic interactions.* Ecology, 88(10), pp.2415-2426.

Sevilla, A. and Glotin, H., (2017). *Audio Bird Classification with Inception-v4 extended with Time and Time-Frequency Attention Mechanisms.* CLEF (Working Notes), 1866, pp.1-8.

Wilder, S.M., Barnes, C.L. and Hawlena, D., (2019). *Predicting predator nutrient intake from prey body contents.* Frontiers in Ecology and Evolution, 7, p.42.

Wootton, J.T. and Emmerson, M. (2005). *Measurement of interaction strength in nature*. Annu. Rev. Ecol. Evol. Syst., 36, pp.419-444.

WWF (2023), *The Effects of Climate Change* [online]. Available at: https://www.wwf.org.uk/learn/effects-of/climate-change [accessed 23/11/2023]

Xambó, A. (2024) 'Week 3: Centrality Measures' [Lecture], *ECS637U: Digital Media and Social Networks*. Queen Mary University, London. 6 February.

Appendix A – Risk Assessment

Description of Risk	Impact of Risk	Likelihood Rating	Impact Rating	Preventative Actions
Training not adequate	Model may not provide accurate results	Low	High	I will ensure to go through multiple training cycles, checking the results each time. I will decide on how many training cycles to carry out based on the accuracy of results.
Results are not an accurate representation of effects of climate change	Project would not be able to correctly analyse effects	Low	High	I will check the results against existing studies about the effects of climate change on the same type of networks to ensure that they are similar or that they follow the same pattern.
Dataset used has a large number of missing values	Less species or feature of species that I will be able to use	High	Low	Although I am unable to prevent this, I will perform data cleaning to ensure all data is in the right state to be analysed accurately.
Unable to access or import data	Unable to perform any analysis on the data	Medium	High	I will ensure that I can access all data needed before attempting to implement it, and I will make necessary adjustments to make data imports as seamless as possible.
Having to pay for certain features of packages I am using	Limited capability in analysis	Low	Medium	I will check the availability of all packages I plan to use and test them, if possible, to ensure that

				I have access to the features I need.
A provided API is discontinued/no longer in service	Unable to carry out tasks originally planned	Low	Medium	Although I am unable to prevent this, I will ensure I use higher tier APIs so that this is less likely to occur. As well as considering other API options which I could use in place of one that is lost.
Having a large number of other assignments or important exams	Spending less time on the project, which could lead to delays in project	Low	Medium	I will keep track of any upcoming deadlines to ensure that I complete all work in good time.
Personal illness/Unforeseen or uncontrolled circumstances	Spending less time or effort on the project, which could lead to delays in project	Medium	Medium	I will inform my supervisor and school if the circumstance is severe and figure out what steps can be taken to ensure I complete the project in the right time frame.

Appendix B – Time Plan

ID	Task Name	Start	Finish	Deadline
1	Project Definition	25/09/2023	12/10/2023	16/10/2023
2	Complete Background Research	10/10/2023	30/10/2023	
3	Learn Cypher/Familiarise with Neo4j	20/10/2023	06/11/2023	
4	Import Dataset	05/11/2023	05/11/2023	
5	Initial Analysis of Dataset	06/11/2023	20/11/2023	
6	Interim Report	16/10/2023	25/11/2023	27/11/2023
7	Progress Presentation	10/11/2023	01/12/2023	04/12/2023
8	Split Database – Filter by Location	05/12/2023	20/12/2023	
9	Build Graph of Networks	16/01/2024	22/01/2024	
10	Apply Algorithms to Small Network	22/01/2024	29/01/2024	
11	Test Accuracy of Results	29/01/2024	03/02/2024	
12	Select Data to Use	05/02/2024	10/02/2024	
13	Apply Algorithms to Chosen Graphs	10/02/2024	26/02/2024	
14	Analyse Results	20/02/2024	15/03/2024	
15	Draft Report	20/01/2024	15/03/2024	18/03/24
16	Finalise Data to Use	06/03/2024	08/03/2024	

17	Apply Algorithms to Final Graphs	15/03/2024	25/03/2024	
18	Analyse Final Results and Derive Conclusion	25/03/2024	25/04/2024	
19	Final Report	20/03/2024	25/04/2024	29/04/2024
20	Project Video	11/03/2024	28/04/2024	01/05/2024

Appendix C – Source Code

C1. Initial Investigation and Processing of Dataset

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer

df = pd.read_csv('/content/283_2_FoodWebDataBase_2018_12_10.csv')

df_clean = df.copy()

# change object columns to category
df_clean[df_clean.select_dtypes(['object']).columns]                    =
df.select_dtypes(['object']).apply(lambda x: x.astype('category'))
display(df_clean.info())

# view sample of dataset
display(df_clean.sample(n=10))

print('Number of missing values:')
df_clean = df_clean.replace(-999.0,np.NaN)

for col in df_clean.columns:
    print('\t%s: %d' % (col,df_clean[col].isna().sum()))

# only include select columns in reduced FoodWebDataBase
df_clean = df_clean[['autoID',
                     'interaction.type',
                     'con.taxonomy',
                     'con.mass.mean.g.',
                     'res.taxonomy',
                     'res.mass.mean.g.',
                     'geographic.location',
                     'longitude',
                     'latitude',
                     'ecosystem.type',
                     'foodweb.name']]
df_clean.to_csv('/content/FoodWebDataBase.csv', index=False)
```

C2. Filtering Dataset

The same method for filtering was used for each food web, with just the longitude value and the name being changed each time. Below is one example of the code that was used.

```
# L1P1 Portugal Marine
mask = df_clean['longitude'].astype(str).str.contains('-9.340457')

df_filtered = df_clean[mask]

df_reduced = df_filtered[['autoID',
                          'interaction.type',
                          'con.taxonomy',
                          'con.mass.mean.g.',
                          'res.taxonomy',
                          'res.mass.mean.g.',
                          'foodweb.name']]
```

```
df_reduced['autoID']=
df_reduced.groupby(['con.taxonomy', 'res.taxonomy']).ngroup()

df_reduced.to_csv('/content/Portugal_L1P1.csv', index=False)
display(df_reduced.sample(5))

# make sure there are no missing values
df_reduced = df_reduced.replace(-999.0,np.NaN)
for col in df_reduced.columns:
    print('\t%s: %d' % (col,df_reduced[col].isna().sum()))

# ensure that there is species overlap
df_l1p1 = pd.read_csv('./Portugal_L1P1.csv')
df_l1p2 = pd.read_csv('./Portugal_L1P2.csv')

df_l1p1['con.overlap']=
df_l1p1['con.taxonomy'].isin(df_l1p2['con.taxonomy'])

df_l1p1['res.overlap']=
df_l1p1['res.taxonomy'].isin(df_l1p2['res.taxonomy'])

display(df_l1p1.sample(10))
```

C3. Creating Graph of Centralities

This is the code that was used to create a graph of the similarity results for each food web.

```
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv('Portugal_Centralities.csv')

# Calculate the error values as the difference between average and
min/max
df['lower_error'] = df['average'] - df['min']
df['upper_error'] = df['max'] - df['average']

# Create asymmetric error values assuming that lower and upper errors
are different
error = [df['lower_error'].values, df['upper_error'].values]

plt.figure(figsize=(10,5))

# Adjust the size of the nodes based on the number of food webs. You may
want to scale the sizes if the numbers are too large or too small.
plt.scatter(df['Species'],   df['average'],   s=df['foodwebs']*10)    #
Multiply by a factor to adjust the size of the nodes

# Add the error bars
plt.errorbar(df['Species'], df['average'], yerr=error, fmt='none')   #
Change 'fmt' to 'none' to remove the line markers

plt.xlabel('Species')
plt.ylabel('Average Degree Centrality')
plt.title('Average Degree Centrality of Species with Error Bars')
plt.xticks(rotation='vertical')
plt.show()
```

C4. Creating Heatmap of Similarity

This is an example of the method that was used to create a heatmap of the similarity results for each food web. The same code was used with different results each time, for the different food webs.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# read csv with similarity scores and adjust data types
df = pd.read_csv('L1P2_similarity.csv')
df = df.astype({'node1':'category', 'node2':'category'})

# define pivot table with two nodes and similarity result
pivot_table= df.pivot(index='node1', columns='node2', values='inverse')

# plot heatmap
sns.heatmap(pivot_table, cmap='coolwarm')
plt.show()
```

This is the code used to create the heatmap for the embedding results.

```
# read csv with similarity scores and adjust data types
df = pd.read_csv('Soc_emb_sim.csv')
df = df.astype({'page1':'category', 'page2':'category'})

# Group by species pairs and calculate the mean similarity
df_mean                    =                    df.groupby(['page1',
'page2'])['similarity'].mean().reset_index()

# define pivot table with two nodes and averaged similarity result
pivot_table     =     df_mean.pivot(index='page1',     columns='page2',
values='similarity')

# plot heatmap
sns.heatmap(pivot_table, cmap='coolwarm')
plt.show()
```

Appendix D – Tables

Degree Centrality Results

Table D1. Top 15 species in L1P1 by degree centrality score

"Lipophrys pholis"	["Consumer"]	33.0
"Gobius paganellus"	["Consumer"]	29.0
"Lipophrys trigloides"	["Consumer"]	27.0
"Saccorhiza polyschides"	["Resource"]	25.0
"Colpomenia peregrina"	["Resource"]	25.0
"Sphacelaria cirrosa"	["Resource"]	25.0
"Asparagopsis armata"	["Resource"]	25.0
"Bifurcaria bifurcata"	["Resource"]	25.0
"Ulva lactuca"	["Resource"]	25.0
"Gelidium corneum"	["Resource"]	25.0
"Diplodus sargus sargus"	["Consumer"]	25.0
"Dictyota dichotoma"	["Resource"]	25.0
"Cladophora rupestris"	["Resource"]	25.0
"Cystoseira sp."	["Resource"]	25.0
"Dictyopteris polypodioides"	["Resource"]	25.0

Table D2. Top 15 species in L1P2 by degree centrality score

"Gobius paganellus"	["Consumer"]	30.0
"Lipophrys pholis"	["Consumer"]	27.0
"Anemonia sulcata"	["Consumer"]	25.0
"Palaemon longirostris"	["Consumer"]	21.0
"Bifurcaria bifurcata"	["Resource"]	20.0
"Dictyopteris polypodioides"	["Resource"]	20.0
"Sphacelaria cirrosa"	["Resource"]	20.0
"Dictyota dichotoma"	["Resource"]	20.0
"Colpomenia peregrina"	["Resource"]	20.0
"Asparagopsis armata"	["Resource"]	20.0
"Palaemon elegans"	["Consumer"]	20.0
"Ulva lactuca"	["Resource"]	20.0
"Ellisolandia elongata"	["Resource"]	20.0
"Zooplankton"	["Resource"]	17.0
"Maja squinado"	["Consumer"]	17.0

Table D3. Top 15 species in L2P1 by degree centrality score

"Gobius paganellus"	["Consumer"]	26.0
"Codium sp."	["Resource"]	25.0
"Sphacelaria cirrosa"	["Resource"]	25.0
"Ellisolandia elongata"	["Resource"]	25.0
"Ceramium ciliatum"	["Resource"]	25.0
"Bifurcaria bifurcata"	["Resource"]	25.0
"Asparagopsis armata"	["Resource"]	25.0
"Ulva lactuca"	["Resource"]	25.0
"Sargassum sp."	["Resource"]	25.0
"Actinia fragacea"	["Consumer"]	24.0
"Lipophrys pholis"	["Consumer"]	24.0
"Palaemon longirostris"	["Consumer"]	21.0
"Palaemon elegans"	["Consumer"]	20.0
"Zooplankton"	["Resource"]	14.0
"Eriphia verrucosa"	["Consumer"]	12.0

Table D4. Top 15 species in CR1P3 by degree centrality score

"Lipophrys pholis"	["Consumer"]	30.0
"Gobius paganellus"	["Consumer"]	30.0
"Leathesia marina"	["Resource"]	22.0
"Chondrus crispus"	["Resource"]	22.0
"Asparagopsis armata"	["Resource"]	22.0
"Sphacelaria cirrosa"	["Resource"]	22.0
"Colpomenia peregrina"	["Resource"]	22.0
"Ellisolandia elongata"	["Resource"]	22.0
"Ulva lactuca"	["Resource"]	22.0
"Codium sp."	["Resource"]	22.0
"Aulactinia verrucosa"	["Consumer"]	19.0
"Actinia equina"	["Consumer"]	19.0
"Zooplankton"	["Resource"]	14.0
"Caprella sp."	["Consumer"]	11.0
"Gammaridae"	["Consumer"]	11.0

Table D5. Top 15 species in CR2P4 by degree centrality score

"Lipophrys pholis"	["Consumer"]	20.0
"Sphacelaria cirrosa"	["Resource"]	19.0
"Ulva lactuca"	["Resource"]	19.0
"Ellisolandia elongata"	["Resource"]	19.0
"Gobius paganellus"	["Consumer"]	19.0
"Pachygrapsus marmoratus"	["Consumer"]	14.0
"Zooplankton"	["Resource"]	14.0
"Palaemon longirostris"	["Consumer"]	13.0
"Carcinus maenas"	["Consumer"]	8.0
"Gammaridae"	["Consumer"]	6.0
"Polychaeta"	["Resource"]	6.0
"PhytoP"	["Resource"]	6.0
"Insecta"	["Consumer"]	5.0
"Insecta"	["Resource"]	5.0
"Ophiuroidea"	["Consumer"]	5.0

Table D6. Degree centrality results for all food webs

Species	Type	L1P1	L1P2	L2P1	CR1P3	CR2P4	average	min	max	foodwebs
"Lipophrys pholis"	["Consumer"]	33	27	24	30	20	26.8	20	33	5
"Sphacelaria cirrosa"	["Resource"]	25	20	25	22	19	22.2	19	25	5
"Ulva lactuca"	["Resource"]	25	20	25	22	19	22.2	19	25	5
"Ellisolandia elongata"	["Resource"]		20	25	22	19	21.5	19	25	4
"Zooplankton"	["Resource"]		17	14	14	14	14.75	14	17	4
"Gobius paganellus"	["Consumer"]	29	30	26	30		28.75	26	30	4
"Asparagopsis armata"	["Resource"]	25	20	25	22		23	20	25	4
"Palaemon longirostris"	["Consumer"]		21	21		13	18.33333	13	21	3
"Colpomenia peregrina"	["Resource"]	25	20		22		22.33333	20	25	3
"Bifurcaria bifurcata"	["Resource"]	25	20	25			23.33333	20	25	3
"Gammaridae"	["Consumer"]				11	6	8.5	6	11	2
"Codium sp."	["Resource"]			25	22		23.5	22	25	2
"Palaemon elegans"	["Consumer"]		20	20			20	20	20	2
"Dictyota dichotoma"	["Resource"]	25	20				22.5	20	25	2
"Dictyopteris polypodioides"	["Resource"]	25	20				22.5	20	25	2
"Gobius paganellus"	["Consumer"]					19	19	19	19	1
"Pachygrapsus marmoratus"	["Consumer"]					14	14	14	14	1
"Carcinus maenas"	["Consumer"]					8	8	8	8	1
"Polychaeta"	["Resource"]					6	6	6	6	1
"PhytoP"	["Resource"]					6	6	6	6	1
"Insecta"	["Resource"]					5	5	5	5	1
"Insecta"	["Consumer"]					5	5	5	5	1
"Ophiuroidea"	["Consumer"]					5	5	5	5	1
"Leathesia marina"	["Resource"]				22		22	22	22	1
"Chondrus crispus"	["Resource"]				22		22	22	22	1
"Aulactinia verrucosa"	["Consumer"]				19		19	19	19	1
"Actinia equina"	["Consumer"]				19		19	19	19	1
"Caprella sp."	["Consumer"]				11		11	11	11	1
"Ceramium ciliatum"	["Resource"]			25			25	25	25	1
"Sargassum sp."	["Resource"]			25			25	25	25	1
"Actinia fragacea"	["Consumer"]			24			24	24	24	1
"Eriphia verrucosa"	["Consumer"]			12			12	12	12	1
"Anemonia sulcata"	["Consumer"]		25				25	25	25	1
"Maja squinado"	["Consumer"]		17				17	17	17	1
"Lipophrys trigloides"	["Consumer"]	27					27	27	27	1
"Saccorhiza polyschides"	["Resource"]	25					25	25	25	1
"Gelidium corneum"	["Resource"]	25					25	25	25	1
"Cladophora rupestris"	["Resource"]	25					25	25	25	1
"Cystoseira sp."	["Resource"]	25					25	25	25	1
"Diplodus sargus sargus"	["Consumer"]	25					25	25	25	1

K-Nearest Neighbours Results

Table D7. Inverse of similarity scores for L1P1 food-web

node1	node2	inverse
"Anemonia sulcata"	"Palaemon elegans"	0.989065
"Anemonia sulcata"	"Lipophrys trigloides"	0.989047
"Anemonia sulcata"	"Lipophrys pholis"	0.989012
"Anemonia sulcata"	"Epitonium clathrus"	0.987911
"Lipophrys pholis"	"Gobius paganellus"	0.312715
"Lipophrys trigloides"	"Gobius paganellus"	0.137931
"Palaemon elegans"	"Diplodus sargus sargus"	0.082569
"Polychaeta"	"Balanus sp."	0.002991
"Polychaeta"	"Chthamalus sp."	0.002991
"Polychaeta"	"Insecta"	0.002991
"Gammaridae"	"Zooplankton"	0.001298
"Pantopoda"	"Saccorhiza polyschides"	0.00127
"Pantopoda"	"Cystoseira sp."	0.00127
"Pantopoda"	"Asparagopsis armata"	0.00127
"Pantopoda"	"Dictyota dichotoma"	0.00127
"Pantopoda"	"Bifurcaria bifurcata"	0.00127
"Pantopoda"	"Dictyopteris polypodioides"	0.00127
"Nucella sp."	"Musculus costulatus"	0.000799

Table D8. Inverse of similarity scores for L1P2 food-web

node1	node2	inverse
"Lipophrys pholis"	"Gobius paganellus"	0.677419
"Maja squinado"	"Gibbula umbilicalis"	0.342105
"Anemonia sulcata"	"Gibbula umbilicalis"	0.236641
"Maja squinado"	"Gobius paganellus"	0.236641
"Anemonia sulcata"	"Palaemon elegans"	0.21875
"Gobius paganellus"	"Acanthochitona crinita"	0.186992
"Anemonia sulcata"	"Maja squinado"	0.173554
"Gobius paganellus"	"Gibbula umbilicalis"	0.173554
"Gobius paganellus"	"Palaemon elegans"	0.152542
"Anemonia sulcata"	"Palaemon longirostris"	0.122807
"Palaemon longirostris"	"Palaemon elegans"	0.122807
"Anemonia sulcata"	"Gobius paganellus"	0.090909
"Palaemon elegans"	"Sphaeromatidae"	0.090909
"Palaemon elegans"	"Ophiuroidea"	0.056604
"Gobius paganellus"	"Palaemon longirostris"	0.038462
"Nucella sp."	"Musculus costulatus"	0.001498
"Patella ulyssiponensis"	"Asparagopsis armata"	0.000971
"Patella ulyssiponensis"	"Colpomenia peregrina"	0.000971
"Patella ulyssiponensis"	"Dictyota dichotoma"	0.000971
"Gammaridae"	"Asparagopsis armata"	0.000771
"Gammaridae"	"Ellisolandia elongata"	0.000771
"Gammaridae"	"Dictyopteris polypodioides"	0.000771
"Omalogyra atomus"	"Asparagopsis armata"	0.000572
"Omalogyra atomus"	"Colpomenia peregrina"	0.000572
"Omalogyra atomus"	"Dictyopteris polypodioides"	0.000572
"Polychaeta"	"Insecta"	0
"Nematoda"	"Oligochaeta"	0
"Ampithoidae"	"Polychaeta"	0

Table D9. Inverse of similarity scores for L2P1 food-web

node1	node2	inverse
"Anemonia sulcata"	"Palaemon elegans"	0.969724
"Anemonia sulcata"	"Aplysia punctata"	0.968915
"Anemonia sulcata"	"Lipophrys pholis"	0.965986
"Actinia fragacea"	"Polychaeta"	0.363057
"Palaemon elegans"	"Polychaeta"	0.346405
"Lipophrys pholis"	"Paracentrotus lividus"	0.342105
"Actinia fragacea"	"Gibbula umbilicalis"	0.324324
"Gobius paganellus"	"Gibbula umbilicalis"	0.082569
"Gobius paganellus"	"Palaemon longirostris"	0.074074
"Gobius paganellus"	"Acanthochitona crinita"	0.047619
"Actinia fragacea"	"Palaemon elegans"	0.038462
"Palaemon longirostris"	"Sphaeromatidae"	0.019608
"Polychaeta"	"Mytilus galloprovincialis"	0.009901
"Turtonia minuta"	"Zooplankton"	0.0006
"Omalogyra atomus"	"Asparagopsis armata"	0.000472
"Omalogyra atomus"	"Ulva lactuca"	0.000472
"Omalogyra atomus"	"Ceramium ciliatum"	0.000472
"Omalogyra atomus"	"Bifurcaria bifurcata"	0.000472
"Omalogyra atomus"	"Codium sp."	0.000472
"Omalogyra atomus"	"Sargassum sp."	0.000472
"Omalogyra atomus"	"Sphacelaria cirrosa"	0.000472
"Nematoda"	"Oligochaeta"	0
"Gobius paganellus"	"Polychaeta"	0

Table D10. Inverse of similarity scores for CR1P3 food-web

node1	node2	inverse
"Anemonia sulcata"	"Mytilus galloprovincialis"	0.957178949
"Anemonia sulcata"	"Gibbula umbilicalis"	0.957028061
"Anemonia sulcata"	"Patella ulyssiponensis"	0.953917051
"Octopus vulgaris"	"Patella depressa"	0.671484888
"Octopus vulgaris"	"Mytilus galloprovincialis"	0.669202779
"Octopus vulgaris"	"Gibbula umbilicalis"	0.659979599
"Octopus vulgaris"	"Patella ulyssiponensis"	0.270072993
"Actinia equina"	"Patella depressa"	0.238964992
"Actinia equina"	"Mytilus galloprovincialis"	0.226604795
"Actinia equina"	"Gibbula umbilicalis"	0.174236168
"Gobius paganellus"	"Patella depressa"	0.103942652
"Lipophrys pholis"	"Mytilus galloprovincialis"	0.088422972
"Lipophrys pholis"	"Patella depressa"	0.07063197
"Lipophrys pholis"	"Gobius paganellus"	0.038461538
"Gobius paganellus"	"Lipophrys pholis"	0.038461538
"Lipophrys pholis"	"Acanthochitona crinita"	0.029126214
"Gobius paganellus"	"Acanthochitona crinita"	0.00990099
"Nematoda"	"Oligochaeta"	0.008133307
"Calliopidae"	"Polychaeta"	0.004182434
"Stenothoidae"	"Colpomenia peregrina"	0.000271926
"Insecta"	"Asparagopsis armata"	0.000271926
"Pantopoda"	"Leathesia marina"	0.000271926
"Omalogyra atomus"	"Sphacelaria cirrosa"	0.00017197
"Omalogyra atomus"	"Leathesia marina"	0.00017197
"Omalogyra atomus"	"Chondrus crispus"	0.00017197
"Gammaridae"	"Zooplankton"	9.999E-05
"Copepoda"	"Zooplankton"	0

Table D11. Inverse of similarity scores for CR2P4 food-web

node1	node2	inverse
"Carcinus maenas"	"Lepidochitona cinerea"	0.990968208
"Carcinus maenas"	"Patella ulyssiponensis"	0.989730951
"Pachygrapsus marmoratus"	"Gibbula umbilicalis"	0.728113105
"Pachygrapsus marmoratus"	"Phorcus lineatus"	0.71807161
"Pachygrapsus marmoratus"	"Lepidochitona cinerea"	0.656475438
"Nassariidae"	"Lipophrys pholis"	0.596774194
"Nassariidae"	"Gobius paganellus"	0.577524292
"Gobius paganellus"	"Acanthochitona crinita"	0.128919861
"Gobius paganellus"	"Leptochiton algesirensis"	0.125109361
"Gobius paganellus"	"Palaemon longirostris"	0.120492524
"Gobius paganellus"	"Lipophrys pholis"	0.101527403
"Lipophrys pholis"	"Sphaeromatidae"	0.058380414
"Palaemon longirostris"	"Sphaeromatidae"	0.036608863
"Lipophrys pholis"	"Acanthochitona crinita"	0.033816425
"Lipophrys pholis"	"Leptochiton algesirensis"	0.029126214
"Polychaeta"	"Mytilus galloprovincialis"	0.004975124
"Insecta"	"Ellisolandia elongata"	0.000471777
"Insecta"	"Sphacelaria cirrosa"	0.000471777
"Insecta"	"Ulva lactuca"	0.000471777
"Pantopoda"	"Ulva lactuca"	0.000471777
"Pantopoda"	"Sphacelaria cirrosa"	0.000471777
"Pantopoda"	"Ellisolandia elongata"	0.000471777
"Tanaidacea"	"Zooplankton"	0.00039984
"Insecta"	"Zooplankton"	0.00029991
"Copepoda"	"Zooplankton"	0
"Gammaridae"	"Oligochaeta"	0

Node Embedding Results

Table D12. Sample of similarity of node embeddings for Facebook network

Page1	*Page2*	*similarity*
"Josh Marks"	"Bobby's Burger Palace"	0.567578793
"Pat Neely"	"SORTEDfood"	0.600471973
"Tom Aikens"	"McDonald's"	0.580545127
"Chef Michelle Bernstein"	"Noëmie Honiat Top Chef"	0.556239188
"Luke Thomas"	"Franklin Becker"	0.573783934
"Nicko's Kitchen"	"Alex French Guy Cooking"	0.586829126
"Blue Ribbon Restaurants"	"Chef Mary Sue Milliken"	0.574688137
"Jose Garces"	"McDonald's"	0.566399574
"T-rex Cafe"	"Dina Nikolaou"	0.617155373